Digitization of library material in Europe
Problems, obstacles and perspectives anno 2007

By Erland Kolding Nielsen

This paper deals only with retro digitization of library material in physical formats, not with digital born material whether this is acquired or harvested by libraries or archives.

Erland Kolding Nielsen,
Director General of the Royal Library,Copenhagen
past President of LIBER
and ex officio member of CENL
ekn@kb.dk

Abstract: *This paper is a short introductory policy paper about the state-of-the art of digitization of library material in Europe, seen from the chief executive point of view of a big national and university library in the autumn of 2007. It focuses on current problems, obstacles, and some perspectives. What has been achieved, what are the problems and obstacles in terms of especially mass digitization in the light of the so-called Google challenge and the response by the Commission of the European Union, and what are the consequences likely to be?*

Introduction

Digitization of library and archive material has been part of library activities for about 15 years. Many national libraries and big university libraries, not to mention archives and other cultural institutions, have digitized – normally smaller – parts of their collections often without an overall plan or at any rate coordinated at national level. National schemes for systematic digitization are very rare, if existing anywhere at all.

The efforts and results are although both vast and many and very difficult to overview. The first documented overview of the digitization within national libraries has been carried out by the National Library of Austria on behalf of *The Conference of European National Librarians* (CENL) this year, and will be presented at this conference. Some of the conclusions are astonishing and rather disturbing.

What are, then, the characteristics of the achievement of libraries during these first 15 years? The CENL investigation reveals that only 1 % of the holdings has been digitized, i.e. approximately 4.7 mill. items, representing 17 mill. pages, so far.

- The main emphasis of digitization has been within newspapers, special collections and rare, fragile or heavily used material within this category, i.e. manuscripts, rare books, photos, maps etc.
- The priority and reason for digitization has been access, not preservation.
- Standards of digitization formats have varied, and are only now in the process of being settled and agreed upon in terms of permanence and preservation.
- Mass digitization has with two exceptions not been planned or carried out.
- Books (and journals), comprising only 12 %, i.e. 619,000 (of which 80 % again are Russian dissertations) have - again with two notable exceptions - not yet been systematically digitized at any systematic national level.
- Hardly any books from the 20th century have been digitized.
- European scholarly journals have not yet been retro digitized in any European country.

The investigation also shows what it will look like in 2012, if the present policies, priorities and financial conditions continue. The coming EDL, *European Digital Library*, will be a library without books!

The case of the Royal Library

As a typical example of what a big – if not one of the biggest in Europe – national and university library like the Royal Library has done since it published the first digital texts on CD-ROM in 1989 and until now, where everything can be accessed over the Internet on the website, managed by a CMS, a *content management system*, and stored in a so-called DOMS, *digital object management system*, the following short overview can be considered representative for the present situation.

We have now digitized c. 175,000 digital objects, varying from fiction books over manuscripts to photographs, often as collections with new names for marketing purposes[1].

- Books: Access to a selection of important Danish literary classics until 1937. Special full text database *The Digital Archive of Danish Literature* (fiction). 2,300 books, 310,000 pp. since 1996.
- *Government reports since 1848*: c. 1,200, c. 200,000 pp. 2005-06.

[1] The digitized collections are described on the URL which gives access to the collections, central URL: www.kb.dk/da/nb/materialer/e-ressourcer/index.html

National schemes for systematic digitization are very rare, if existing anywhere at all.

- Manuscripts, archives, and rare books: Access to a small selection of important manuscripts and early prints (mostly Danish and European). 350 mss. and rare books, 71,000 pp. since 1996.
- Music: Access to a selection of musical scores, incl. manuscripts, mostly Danish composers. 3,600 prints and mss. 106,000 pp. since 1996.
- Photographs and maps: Access to a selection from the Danish National Photo Archive and the map collection. Subjects: Danish topography and portraits. 148,500 items digitised since 1996.
- Serials: Access to a small selection of Danish journals. A Danish counterpart to JStore is published this autumn with full retro digitisation of the first 10 main journals (ca. 35,000 articles) from the 19th century until today, called "tidsskrift.dk" (journal.dk). 14 serials digitised 316,000 pp. since 2003.

This overview is probably more or less representative of the current situation of many European national libraries anno 2007. It is apparently not enough, and it certainly does not address the problems of mass digitization of books and journals, the core collections of national and university libraries.

What are the present obstacles?

What are the obstacles for speeding up this situation and providing more digital content in the years to come?

Technology

The technology of digitization has developed fast over the last 5-8 years, and today I do not think that technology poses a problem, perhaps except a financial one. Scanners for different types of material, size and condition have been developed, including those that are necessary for valuable and/or fragile material. A bigger problem seems in many institutions and countries to be an organizational one: how to organize the digitization business – the production flow, most effectively and efficiently, especially on a broader scale regionally or nationally.

Digitization formats and preservation

The first 15 years showed a range of different formats, some of which were not liable or sufficient in terms of preservation, and that means that the digital material cannot always survive in new digital surroundings without enormous cost of preservation. It is now clear that at least parts of what have already been digitized have to be digitized yet again in order to secure that output meets the requirements of current e-publishing and preservation standards.

At present, the cost of the full digitising process is still very high, and we can hardly imagine the process repeated even though technological advances make this desirable. This indicates that applicable standards for digitizing must support a compromise between the two extremes in financial terms:

1) Digitising for access on de facto browsers/ players and
2) Digitising for substitution.

It is important from the perspective of a European Digital Library that the libraries can agree on formats suitable as a basis for access formats as well as long term preservation formats. Focusing on a rather limited number of open formats combined with strong collaboration within the library world should make it possible to define a dynamic set of best practices safeguarding the investment as long as possible.

Finance

Most national and some university libraries have already redirected quite large resources to digitization purposes, but as they rarely have got sufficient money for their overall activities, it is impossible to finance really big programs, e.g. mass digitization of books and journals. There is only one exception from this (France), and perhaps one or two under way, but not clarified yet.

Organization

In Denmark, we have an ongoing debate on who shall digitize how much for how many. We have this year established a business case that shows that the cheapest way of organizing digitization on a big scale is to concentrate the process and build up advanced digitization competence in a few big institutions.

The situation seems to be similar in other countries. Too many - often small - institutions or institutions with relevant collections of too modest volume want to digitize too little at too high a cost without being able to justify distributed costs of investment and management.

Copyright

By far the biggest obstacle today to digitization of material even after 1880 – apart from the financing – is the present legal situation of European copyright and the conditions and possibilities of negotiating and acquiring the right to digitize objects within the 70 years' limit of the death of the copyright holder.

The extension of the copyright limit from 50 to 70 years after the death of the copyright holder was simply a catastrophe and an enormous obstacle to developing a relevant, adequate and comprehensive EDL with 20th century material of sufficient importance. The frequently emphasized balance between the interests of the copyright holders and the users, in casu the institutions trying to convert the physical material into the digital, has completely tilted to the advantage of the copyright holders. The legal demands of investigating and finding the heirs etc. are simply prohibitive for mass digitization projects with contents from the 20th century.

The sooner the European Commission understands this and acts accordingly, the better the chances of developing a comprehensive and relevant EDL at a level of cost of both production and administration within the range and possibilities of the institutions in question.

The Google challenge

The announcement in December 2004 by Google that they would start a massive digitization program of books – digitizing "the world's knowledge" (15 mill. books from originally six major research libraries) based on entire university library holdings from the 19th and 20th centuries especially from the USA and UK[2] was considered an enormous challenge to almost all European countries, as it could be foreseen that only fragments – and even arbitrary parts – of the national imprint would be incorporated. And the consequence of that could be that Anglo-American books would in the future dominate at all levels of education, research, scholarship and public use. I did agree then – and still do – with most of the main points of criticism voiced by my former French colleague, Jean-Noël Jeanneney in his book *Google and the Myth of Universal Knowledge: A View from Europe*[3].

The response of the European Union

The response came quickly, but – in my opinion - inadequately, from the European Union on

[2]Cf. Ronald Milne: *"The Google Mass Digitization Project at Oxford"*, LIBER Quarterly, vol. 16: 3-4, 2006.

[3]French ed. April 2006, Eng. s.y. Cf. also David Bearman: *"Jean-Noël Jeanneney's Critique of Google: Private Sector Book Digitization and Digital Library Policy",* D-Lib Magazine, December 2006, vol. 12:12.

It is important from the perspective of a European Digital Library that the libraries can agree on formats suitable as a basis for access formats as well as long term preservation formats.

The problem of mass digitization might be stated this way: it is either the state (the public sector) or Google! So what do we want?

Ligue des Bibliothèques Européennes de Recherche
Association of european Research Libraries

September 30, 2005, with the communication called *I2010: Digital Libraries*[4], followed up by an extensive hearing process within the library, archive, museum and cultural sectors of Europe[5], and finally the *Recommendation on the digitisation and online accessibility of cultural material and digital preservation* by the Commission of August 24, 2006, to which all ministers of culture agreed in November 2006. This is the framework for digitisation policy actions of the European Commission in the years to come, including the project of the EDL, acronym for *The European Digital Library*, based on the already established service TEL, *The European Library*, a portal introduced by the Conference of European National Librarians some years ago.

I assume that you are all aware of the vision and content of the communication. There is, of course, in my opinion nothing wrong with the vision: a *European Digital Library* with more than 12 mill. objects by 2012, but the way the Commission addresses the financial problem of mass digitization and especially its expectations as to the possible results of public-private partnerships (PPP) are unrealistic, as there is no market in most European countries for digital products of this kind, with the exception of the English and Spanish speaking world, not even the French.

Conclusions

At the first presentation of the CENL-survey a month ago, it was concluded: *"On institutional level systematic content digitisation is daily practice in many European National Libraries. On the national and EU level there is a need for co-ordinated funding of mass digitization and building up a digital library infrastructure"*.

Today we can foresee that the European Union will reach its goal in terms of digital content, defined as expected number of digital objects, in the EDL, even if the present situation, priorities, and level of activity should continue, but as it will be shown later today not only the National Libraries, the member states and the Union will have to address the problem that this constitutes a great risk! It can be predicted, too, that if priorities in financing and resource allocations are not changed – the EDL in 2010 or 2012 will still consist of mostly digital heritage objects (which is of course in itself not bad at all), but of all other categories than books and journals.

Why is that? Well, simply because if governments refuse to pay for mass digitization of their national imprint of books and journals, this will either not be done or done alone by private firms on terms that we normally are not willingly to accept in Europe.

Accordingly, the political issue to be addressed at both European and national level in all 47 European countries is: Who is going to pay for mass digitization of books and journals, and what restrictions to public or general use shall, will or must we accept in the future if it's done entirely by the private sector, i.e. Google or Microsoft?

A European digital library without access to the most important parts of the past and present from most scholarly angles – books and journals – is a chimaera even in the age of the Internet. But it is a prediction that might be fulfilled in due course, simply because the governments are too slow or do not see the threat, and because the Commission has really not understood the urgency of the problem, as it emerges today.

The problem of mass digitization might be stated this way: it is either the state (the public sector) or Google! So what do we want? Free access or restricted access to what has been free so far in the physical world, but now on market terms in a marketplace without real competition?

I hope that this conference shall address this problem among many others. Thank you for your attention.

[4] http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm. The website has a good overview of the policy actions and documents within the field.
[5] Responses from LIBER and CENL cf. their websites.

Accordingly, the political issue to be addressed at both European and national level in all 47 European countries is: Who is going to pay for mass digitization of books and journals, and what restrictions to public or general use shall, will or must we accept in the future if it's done entirely by the private sector, i.e. Google or Microsoft?